Meta-NanoSim: metagenome simulator for nanopore reads

Theodora Lo[1,2,*], Chen Yang[1,2,*], Ka Ming Nip [1,2], Saber Hafezqorani[1,2], René L Warren[1], Inanc Birol[1,2]

1 Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada
2 University of British Columbia, Vancouver, BC, Canada
* The authors contributed equally

http://www.birollab.ca/

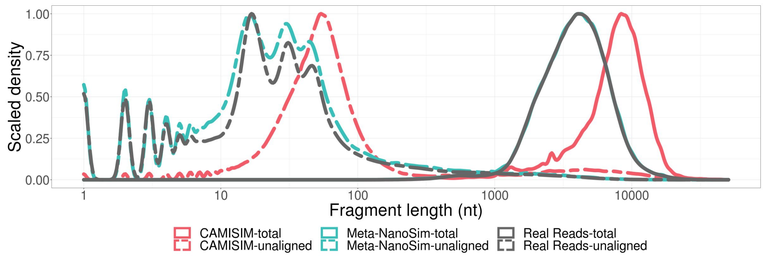
Introduction

- Metagenomic studies using Oxford Nanopore sequencing (ONT) are increasing in frequency due to long read lengths
- High error rates and non-uniform error distributions necessitate the development of tools specific to long reads
- Simulated data with known ground truth can facilitate development of such tools

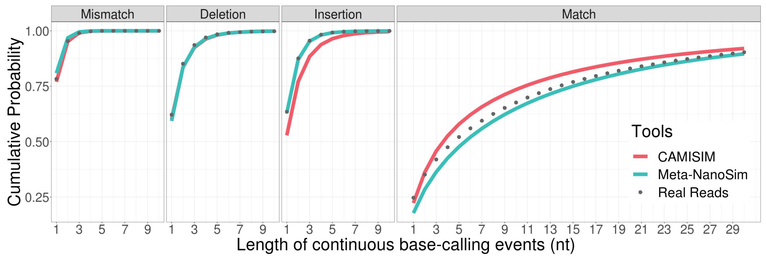
Results

- Benchmarked using ZymoBiomics mock microbial standard consisting of eight bacterial and two yeast species distributed on a log scale (ENA ERR3152366)

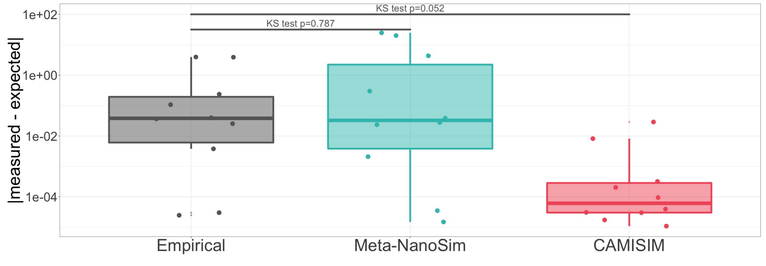
Read length distribution



Probability of consecutive errors/matches



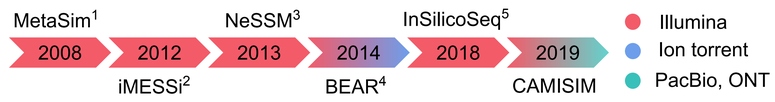
Relative abundances



Conclusion

Meta-NanoSim can simulate microbial communities, consisting of both circular and linear genomes, at user-specified abundances.

Existing metagenome simulators

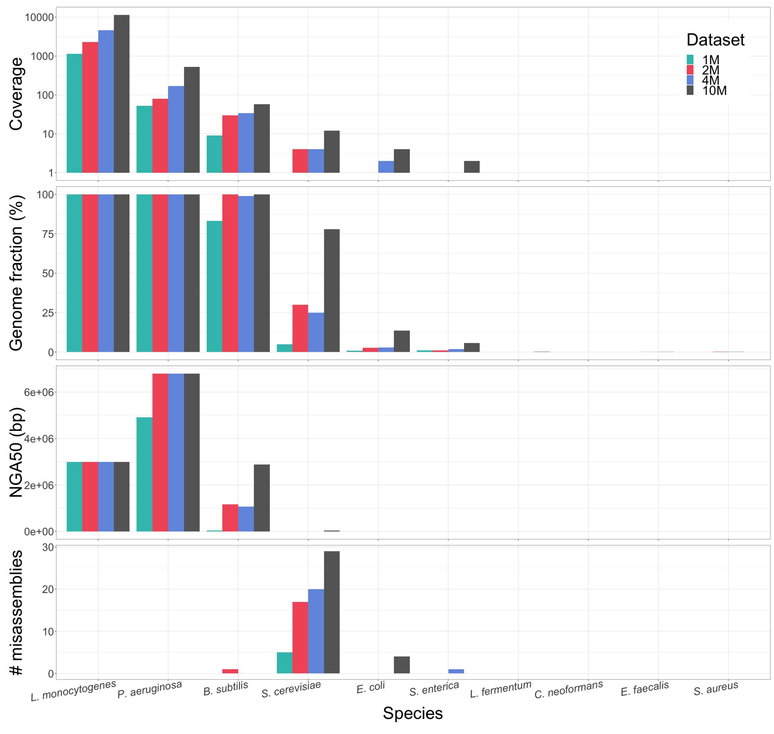


- Only CAMISIM (PMID: 30736849) can simulate ONT metagenome reads, but relies on NanoSim (PMID: 28327957)

1. PMID: 18841204, 2. PMID: 22384016, 3. PMID: 24124490, 4. 25253095, 5. PMID: 30016412

Application: Metagenome assembler benchmark

- Trained Meta-NanoSim on the log microbial community and simulated several datasets with different number of reads
- Assembled with metaFlye (DOI: 10.1101/637637) and evaluated with MetaQUAST (PMID: 26614127)



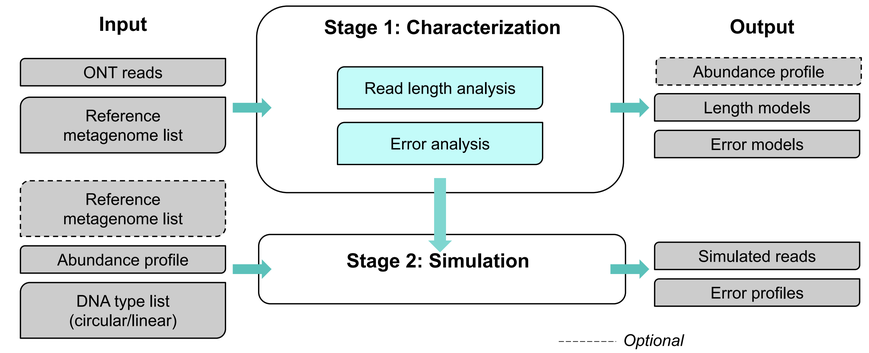
Funding

Genome BritishColumbia
GenomeCanada
iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Github

https://github.com/bcgsc/NanoSim

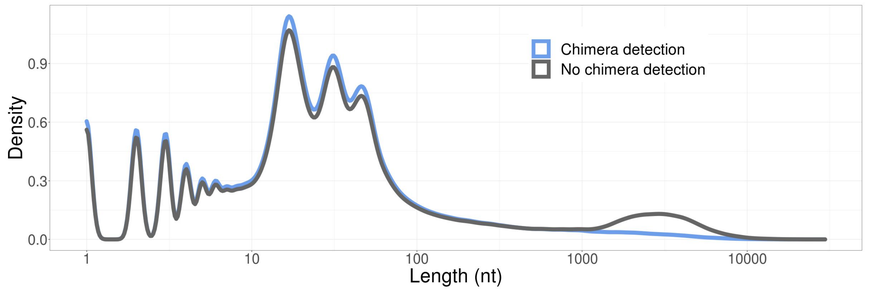
Meta-NanoSim workflow



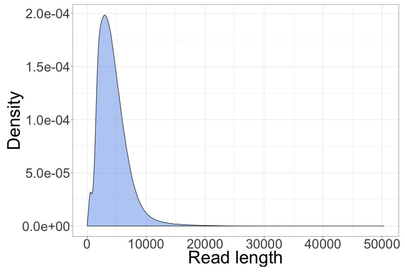
Stage 1: Characterization

Read length distributions and error models are determined from primary and compatible supplementary alignments.

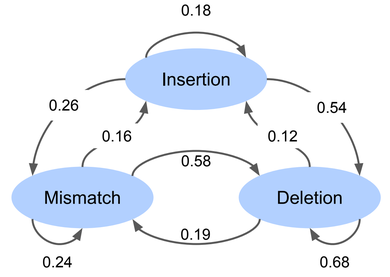
Detect reads that align to circular genomes/plasmids and chimeric reads to ensure accurate length models (optional)



Kernel density estimation used to model read length



Markov chain used to model the probability of each error type



Stage 2: Simulation

Given an abundance profile and DNA type list, can simulate single and multi-sample datasets of complex microbial communities.